Chapter 14
Why Match? Matched Case-Control Studies

Sherri Rose, Mark J. van der Laan

Individually matched case-control study designs are common in public health and medicine, and conditional logistic regression in a parametric statistical model is the tool most commonly used to analyze these studies. In an individually matched case-control study, the population of interest is identified, and cases are randomly sampled. Each of these cases is then matched to one or more controls based on a variable (or variables) believed to be a confounder. The main potential benefit of matching in case-control studies is a gain in efficiency, not the elimination of confounding. Therefore, when are these study designs truly beneficial?

Given the potential drawbacks, including extra cost, added time for enrollment, increased bias, and potential loss in efficiency, the use of matching in case-control study designs warrants careful evaluation.

In this chapter, we focus on individual matching in case-control studies where the researcher is interested in estimating a causal effect, and certain prevalence probabilities are known or estimated. In order to eliminate the bias caused by the matched case-control sampling design, this technique relies on knowledge of the true prevalence probability \( q_0 \equiv P_{X,0}(Y = 1) \) and an additional value:

\[
\bar{q}_0(M) \equiv q_0 \frac{P_{X,0}(Y = 0 \mid M)}{P_{X,0}(Y = 1 \mid M)},
\]

where \( M \) is the matching variable. We will compare the use of CCW-TMLEs in matched and unmatched case-control study designs as we explore which design yields the most information for the causal effect of interest. We assume readers have knowledge of the information presented in the previous chapter on independent case-control study designs.
14.1 Data, Model, and Target Parameter

We define \( X = (W, M, A, Y) \sim P_{X,0} \) as the experimental unit and corresponding distribution \( P_{X,0} \) of interest. Here \( X \) consists of baseline covariates \( W \), an exposure variable \( A \), and a binary outcome \( Y \), which defines case or control status. We can define \( \psi^F_0 = \Psi^F(\mathcal{P}_{X,0}) \in \mathbb{R}^d \) of \( \mathcal{P}_{X,0} \) as the causal effect parameter, and for binary exposure \( A \in \{0, 1\} \) we define the risk difference, relative risk, and odds ratio as in the previous chapter. The observed data structure in matched case-control sampling is defined by

\[
O = \left( (M_1, W_1, A_1), (M^j_0 = M_1, W^j_0, A^j_0 : j = 1, \ldots, J) \right) \sim P_0, \text{ with }

(M_1, W_1, A_1) \sim (M, W, A \mid Y = 1) \text{ for cases and }

(M^j_0, W^j_0, A^j_0) \sim (M, W, A \mid Y = 0, M = M_1) \text{ for controls.}
\]

Here \( M \subset W \), and \( M \) is a categorical matching variable. The sampling distribution of data structure \( O \) is described as above with \( P_0 \). Thus, the matched case-control data set contains \( n \) independent and identically distributed observations \( O_1, \ldots, O_n \) with sampling distribution \( P_0 \). The cluster containing one case and the \( J \) controls is the experimental unit, and the marginal distribution of the cluster is specified by the population distribution \( P_{X,0} \). The model \( \mathcal{M}^F \), which possibly includes knowledge of \( q_0 \) or \( \bar{q}_0(M) \), then implies models for the probability distribution of \( O \) consisting of cases \((M_1, W_1, A_1)\) and controls \((M^j_0, W^j_0, A^j_0), j = 1, \ldots, J\).

14.2 CCW-TMLE for Individual Matching

CCW-TMLEs for individually matched case-control studies incorporate knowledge of \( q_0 \) and \( \bar{q}_0(M) \), where \( \bar{q}_0(M) \) is defined as

\[
\bar{q}_0(M) \equiv q_0 \frac{P_{X,0}(Y = 0 \mid M)}{P_{X,0}(Y = 1 \mid M)} = q_0 \frac{q_0(0 \mid M)}{q_0(1 \mid M)}.
\]

Implementation of CCW-TMLE in individually matched studies echos the procedure for independent (unmatched) case-control studies, with the exception that the weights now differ. We summarize this procedure assuming the reader is already familiar with the material in the previous chapter. We focus on the risk difference \( \psi^F_{RD,0} = E_{X,0}[E_{X,0}(Y \mid A = 1, W) - E_{X,0}(Y \mid A = 0, W)] \) as an illustrative example.

---

**Implementing CCW-TMLE for Individually Matched Data**

**Step 0.** Assign weights \( q_0 \) to cases and \( \bar{q}_0(M)/J \) to the corresponding \( J \) controls.